Julia Joy & Oscar Fleet
Problem:

We plan on creating a document classification model that accomplishes one of the following two goals (we have not yet settled, but they are quite similar): either classifying the genre of a document, or identifying the author of a given document. This problem is related to course, as it will require the use of a NLP classification model that we have covered in class.

For author classification specifically:

What is interesting about this problem, is that the bag-of-words classification will likely not work. When it comes to genre classification of a document, one can likely classify a document's domain correctly due to domain-specific vernacular. However, with the author classification problem, we will likely need to use context classification in order to analyze the writing style of a given author. We are interested in seeing whether a writing style can be classified due to concrete word-sequences, or whether a model will detect artifacts that are non apparent to a human reader. This is the novel component of our project, in the context of our class, trying to classify *style* as opposed to the meaning or sentiment of a text.

For genre classification specifically:

Delving into genre classification from a short summary of each book is quite interesting because although the bag-of-words classification could likely work, our model will need to learn incredibly domain-specific vernacular to get higher accuracy. Not only does it require parsing the nuances of language, but it also requires an understanding of the underlying themes, motifs, and tropes that define each genre. It could potentially offer a fresh perspective on the structural and thematic elements that define genres, revealing underlying patterns that could be less obvious at a glance. This is the novel component of our project, trying to determine if there exists any interesting underlying patterns that define and separate genres from each other, other than the blatant and expected.

The data we will use:

AC: Datasets including documents that have passed into the public domain (We actually don't know whether using copyrighted material for data training in this way is protected by the fair use doctrine), that cover works of at least 10-20 authors. Experiment with short stories and full novels alike. Use documents that are not in the same 'story' (i.e., have different characters and settings).

GC: Websites such as Goodreads or Amazon Kindle can offer a good source for web scraping to obtain short summaries of novels as well as their corresponding genre. Data will need to be processed and cleaned afterwards to be put in the format needed for our model. Ideally we would be able to build and text our model on a smaller training set (of potentially 50-100 books), and evaluate it on a larger test set (100-200 books). A realistic scale of the data will be better understood, tested and adjusted as data is collected and the model implemented.

Evaluation of results:

Since this is a classification task, our evaluation of results would largely rely on an accuracy score- the percentage of time that our model classifies correctly (either the author or the genre). We would examine rates of precision and recall as well, potentially observing if the predictions vary in correctness based on specific genres/authors.

Week by week timeline for May 21st Completion:

April 8th -14th
- Project proposal in

April 15th - 21st
- Dataset found or collected, data preprocessing

April 22nd - 28th
- Data preprocessing done, building model

April 19th - 5th
- Model building tweaks, attempting to increase the accuracy score if needed

May 6th - 12th
- Model ran on test set, evaluate results

May 13th - 19th
- Finish evaluation of results, project writeup

May 20th - 21st
- Final reread/review and turn in!

Following this schedule outline, our intended deliverable of project check in 2 on May 2nd, would be a cleaned and processed dataset, and a complete (or mostly complete) model along with its corresponding accuracy score.